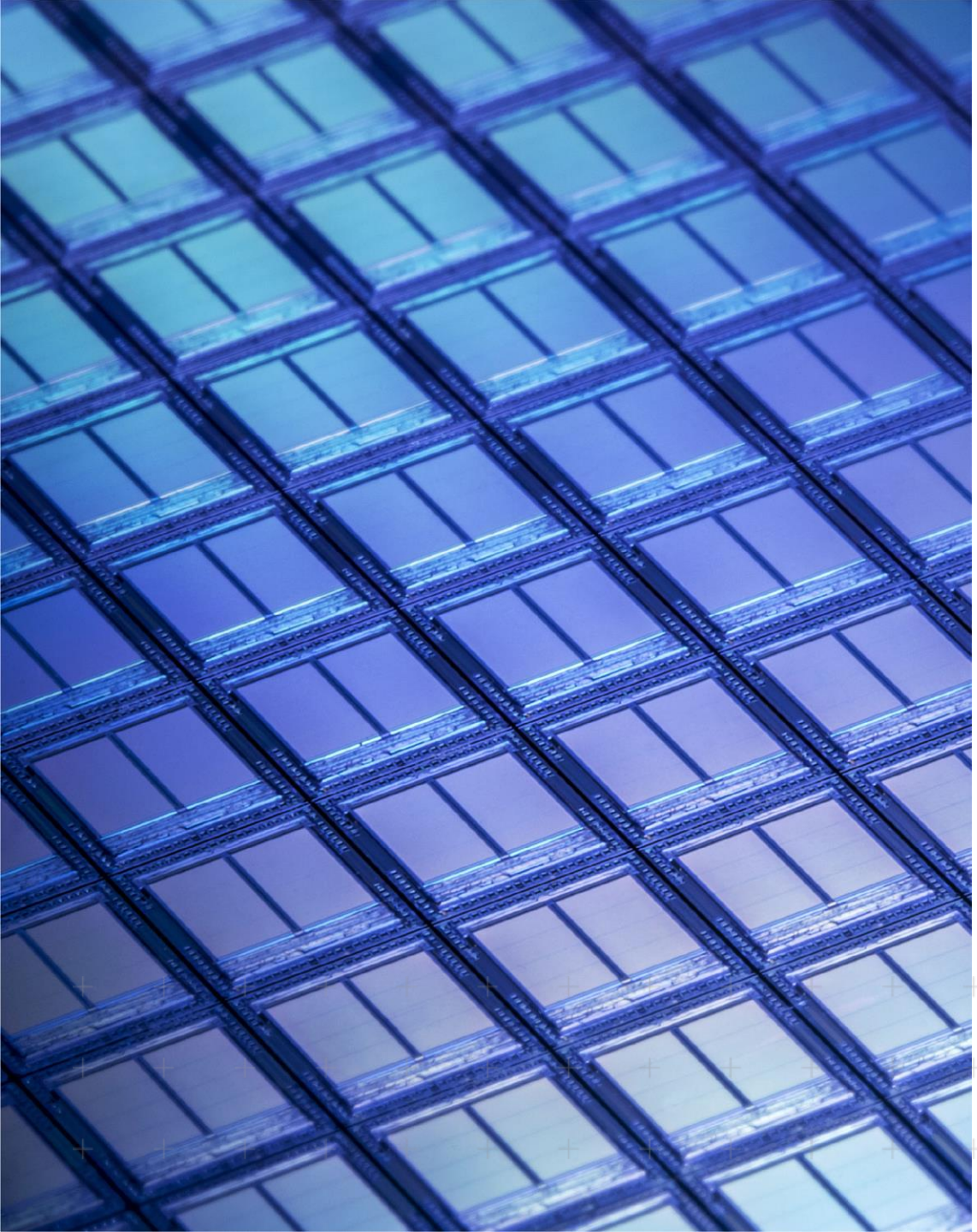


Enabling Advanced Semiconductor High-Volume Manufacturing with Deep Learning

Arvind Jayaraman

AI Algorithm Engineer

KLA Ann Arbor

@ Stanford EE392b

April, 2025

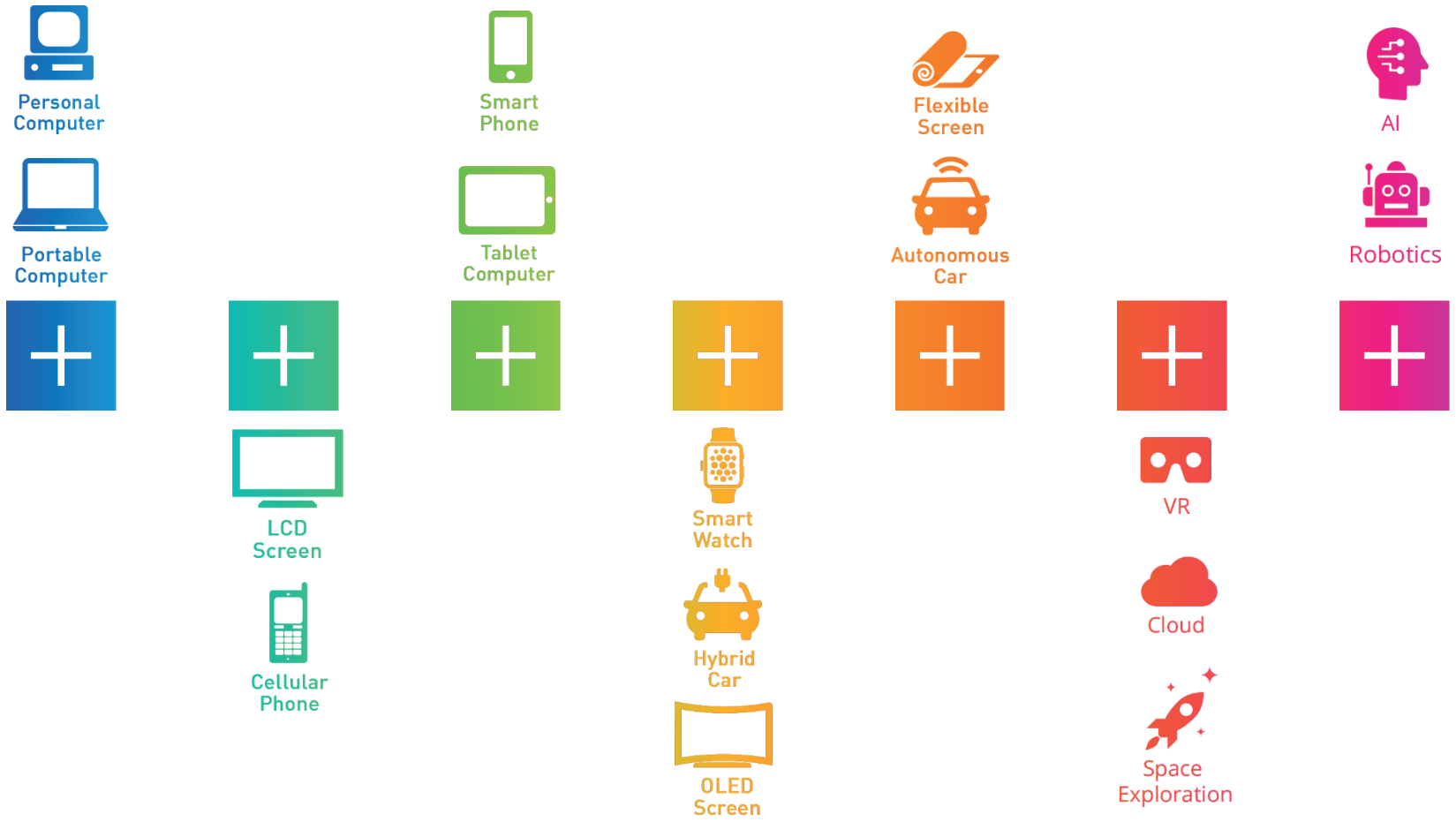
About Me: Arvind Jayaraman

- **Year joined KLA:** 2019
- **Education:** Master's Degree in EE:Systems Univ. Of Michigan, Ann Arbor.
- **Career path:** Algo Engineer in Metrology division to now leading a team that enables Rapid Prototyping of AI & HPC Algos for Optical Inspection Tools.
- **What I work on:** Optical inspection tools
- **Why I chose KLA:**
  - Innovation in bringing products to market that leverage AI
  - KLA is a market leader and a great place to work!

Virtually every electronic device in the world

Is produced using KLA's technologies

KLA's Process Control Tools Power the Chip Industry Today!

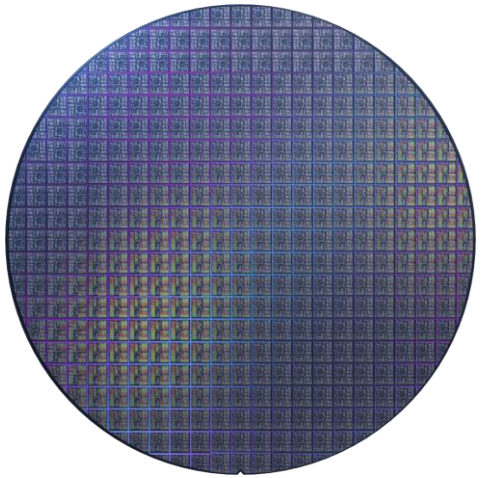


Today's Talk

- What is semiconductor process control?
- Big data to see tiny things
- Artificial Intelligence: not hype, for real!
- High Performance Computing: Making Rubber Meet the Road

What is semiconductor process control?

What is a Chip?

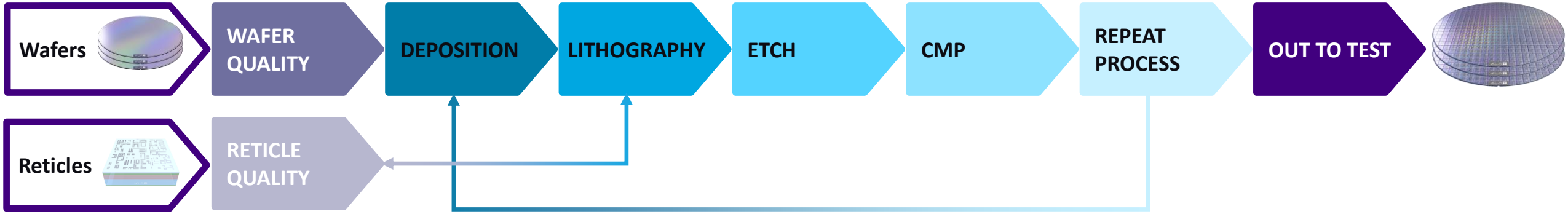


Silicon Wafer

Chip

Chip in a Package

How is a Chip Manufactured?



>1000

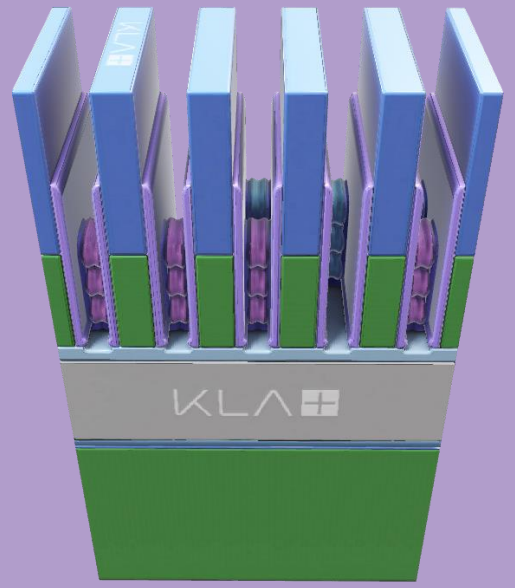
Process Steps

3-6 months

From Bare Wafer → Electrical Test

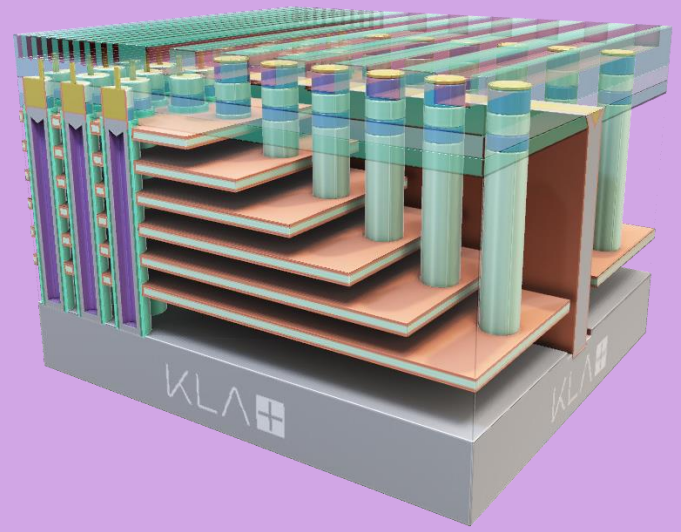
The Manufactured Devices have Complex Geometries

Gate All Around



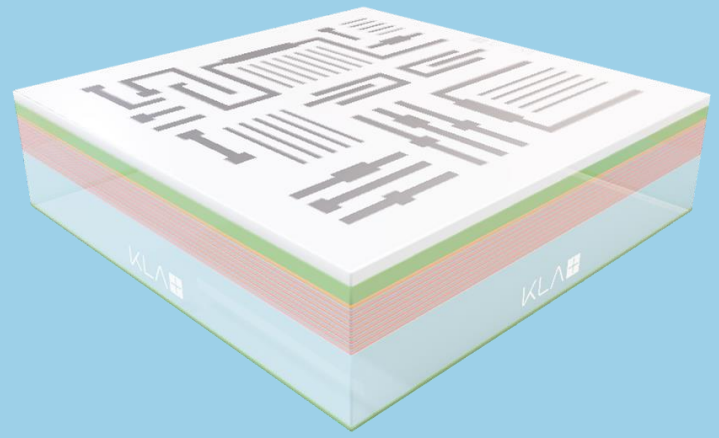
- Complex device features, film stacks
- More process steps, variability

3D NAND



- High aspect ratio (HAR) structure
- Etch and dep variability

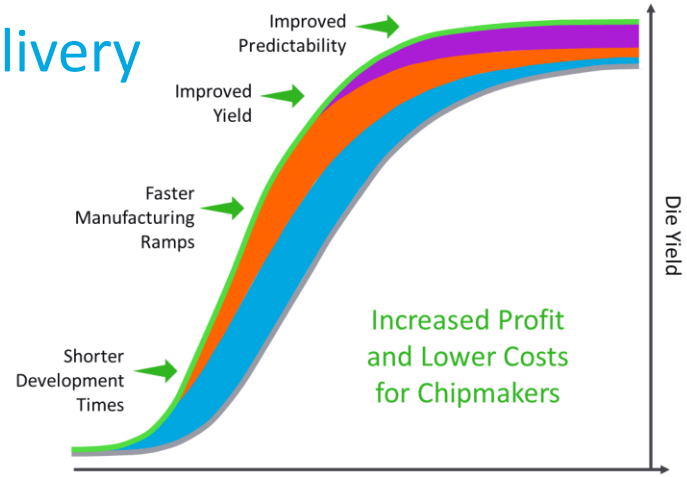
EUV Lithography

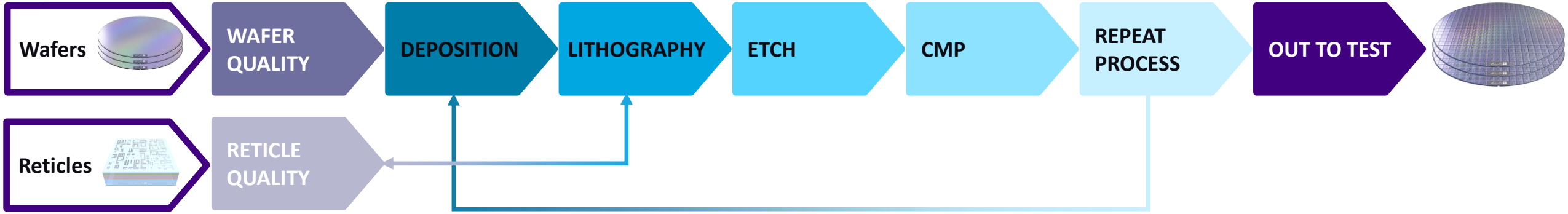


- Complex, multi-layer mask blank
- Small, complex pattern features

KLA Tools Monitor Semiconductor Manufacturing

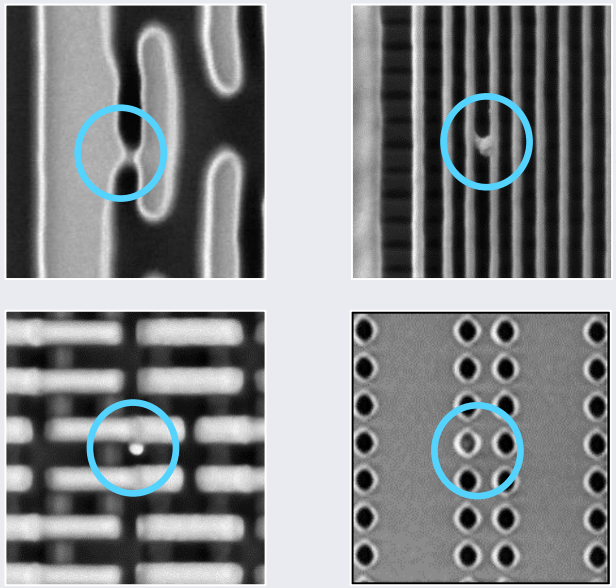
Process Control Enables Faster Yield Ramp & Predictable Product Delivery





We Create the Most Advanced Process Control Systems in the World

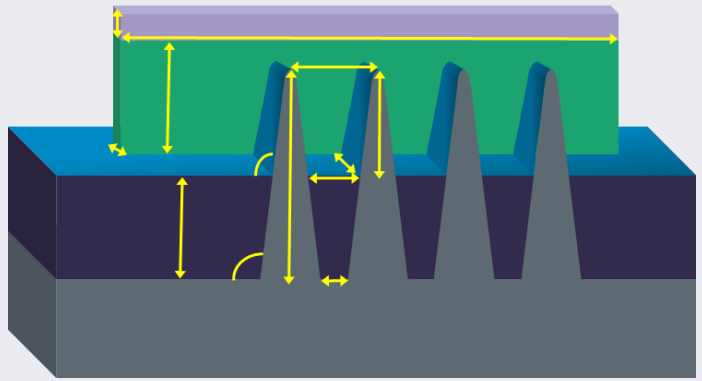
Inspection
Find Critical Defects



Statistically: Outlier Detection

1995: First ever classification system (KLA 2135)
2018: First ever physics-based DL system (KLA eSL10™)

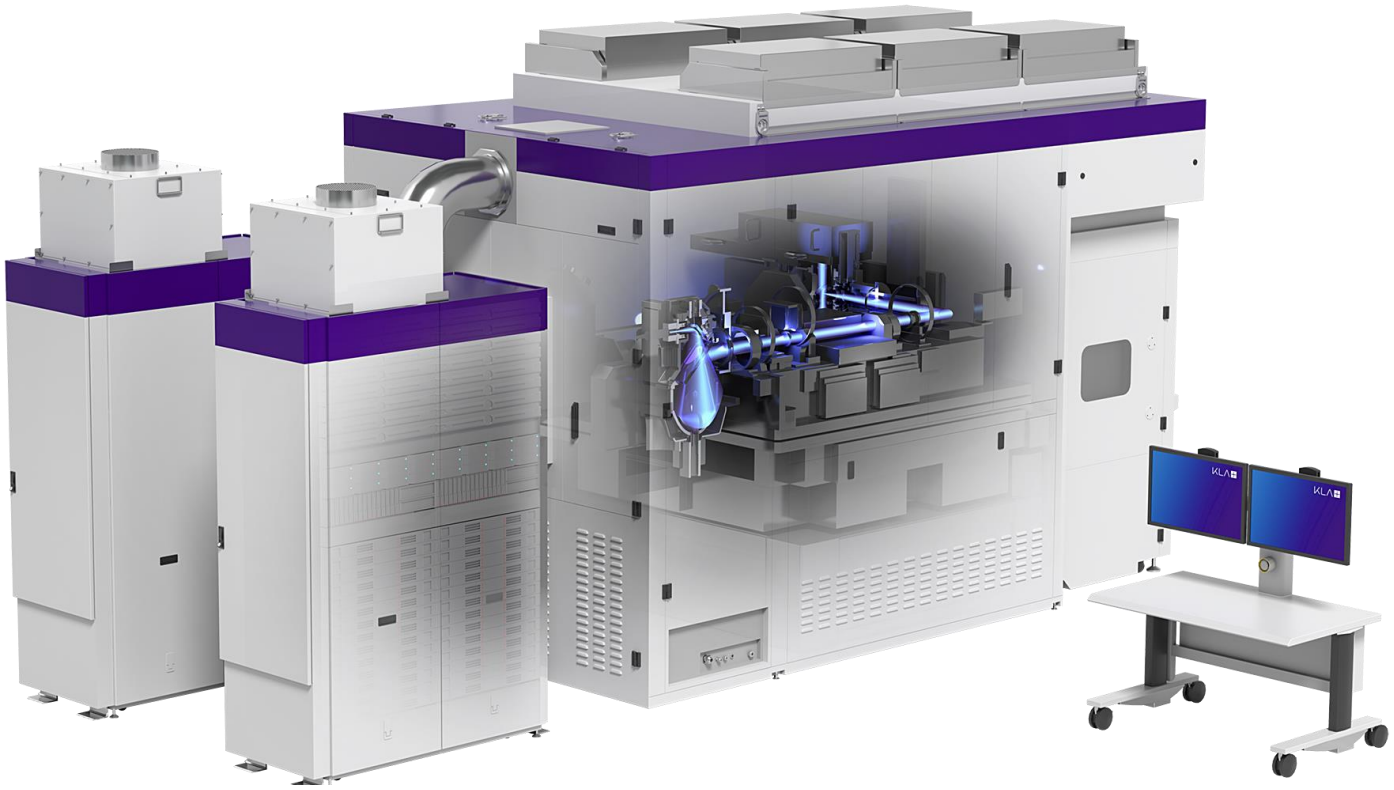
Metrology
Measure Critical Parameters



Statistically: Measure $\mathcal{N}$ (μ , σ^2)

1993: First ever NN-based metrology system (KLA Films)
2017+: Models enhanced by DL

Core Technologies and Expertise



Illumination sources

broadband plasmas, lasers, LEDs, X-rays, electron-beams

Optics

objectives, lenses, mirrors, polarizers, filters for DUV/UV/Vis/IR light, X-rays and electron-beams

Sensors

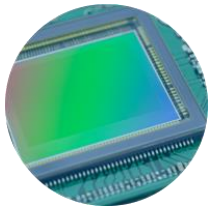
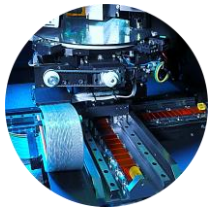


image sensors, photo multiplier tubes, CMOS sensors, cameras

Mechanics

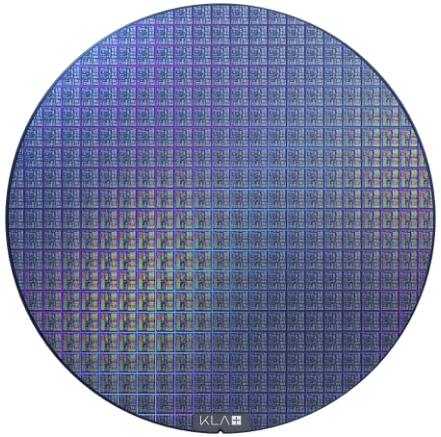


precision stages, motion control, robotics

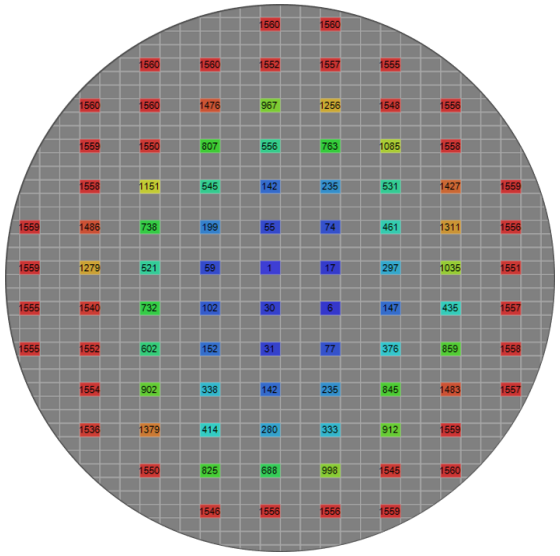
Image and data processing

high-speed data processing, high performance computing, AI/ML/DL, algorithms, computational physics

A KLA Tool in Action



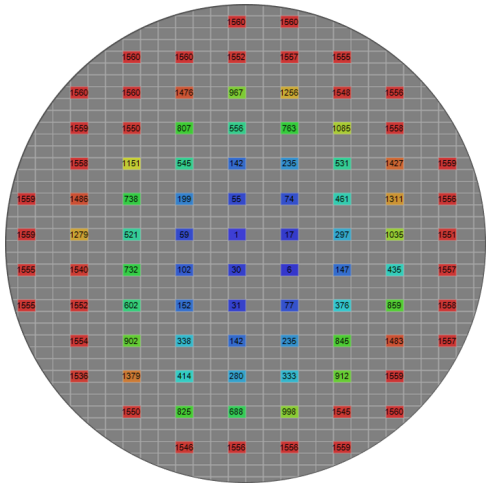
A KLA Tool in Action



10 Billion

number of pixels collected in a 100µm² area with a single scan using Yellowstone™ mode

A KLA Tool in Action



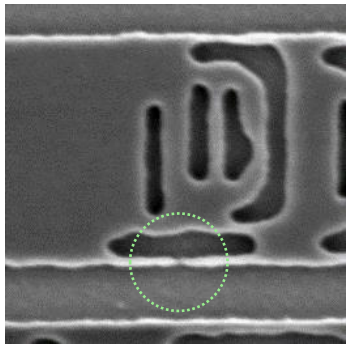


Image: KLA

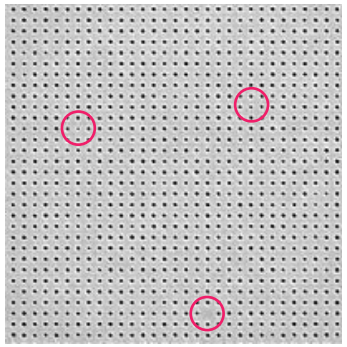


Image: Litho Workshop 2019, imec and KLA

<10NM

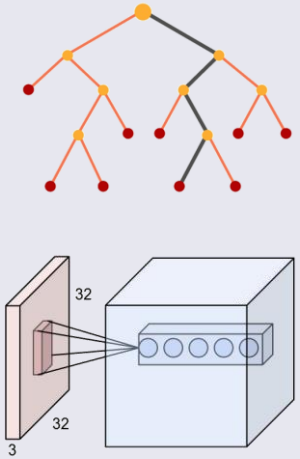
size of the defects that are detected

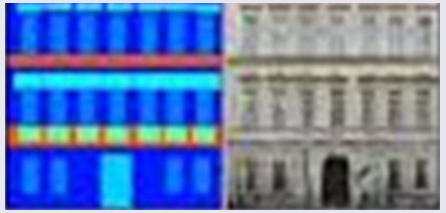
Artificial Intelligence: not hype, for real!

Our Software and Algorithms

Classification

- Random Forests
- Boosted Decision Trees
- MLPs
- CNNs



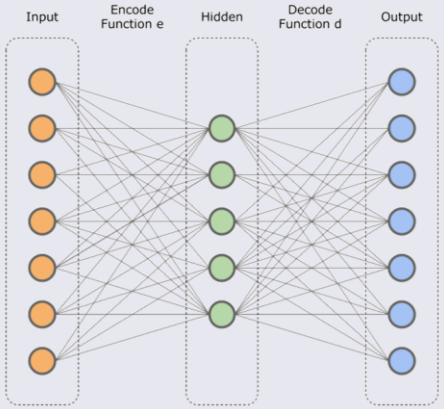


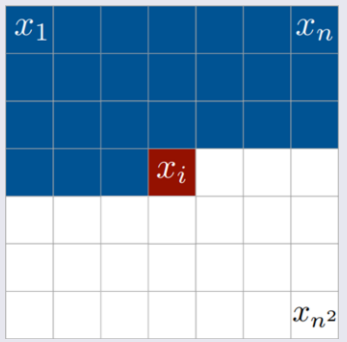
Reference Generation

- Conditional GANs
- VAEs

Natural Grouping and Clustering

- Auto encoders
- Hand crafted features





Active Research Areas

- Physics-based ML
- Pixel CNN

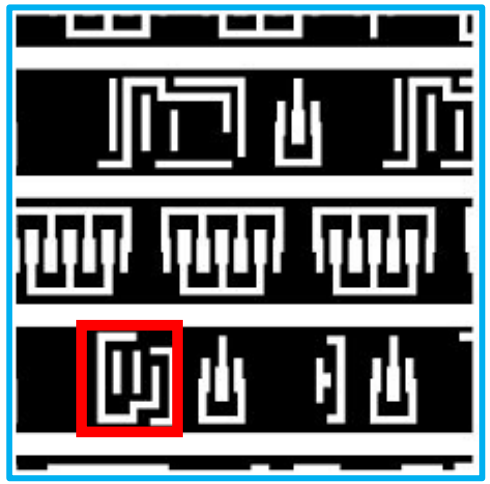
Reference Generation using GANs

Sample of Active DL R&D at KLA

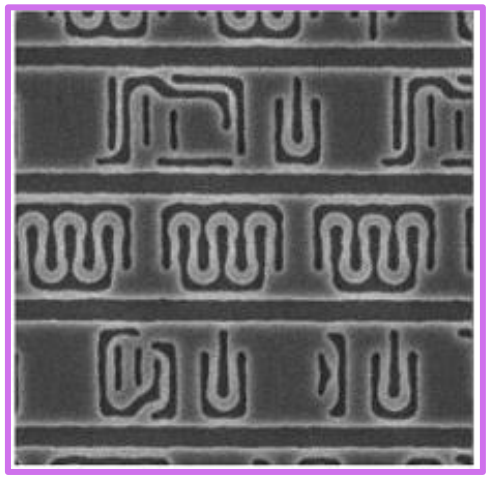
Challenge: Inspect Patch for Defects wrt Design

Defect Patch

Circuit Design



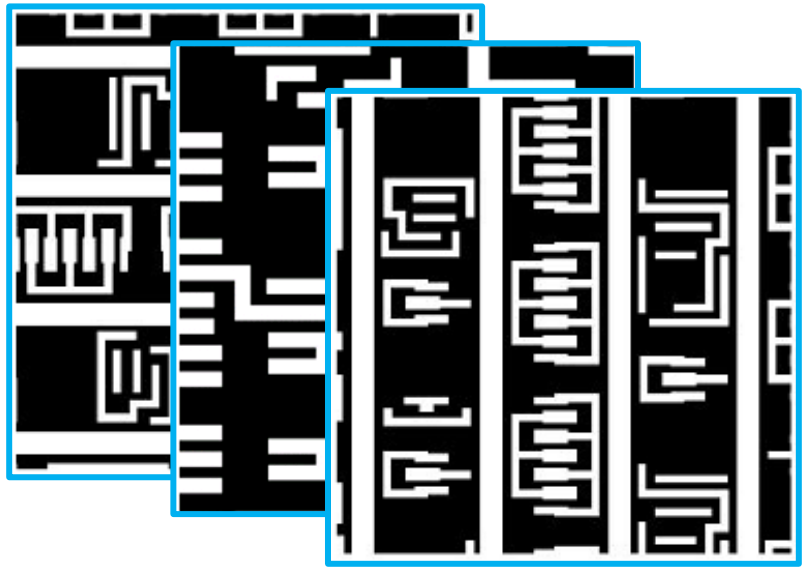
Defect-free Image



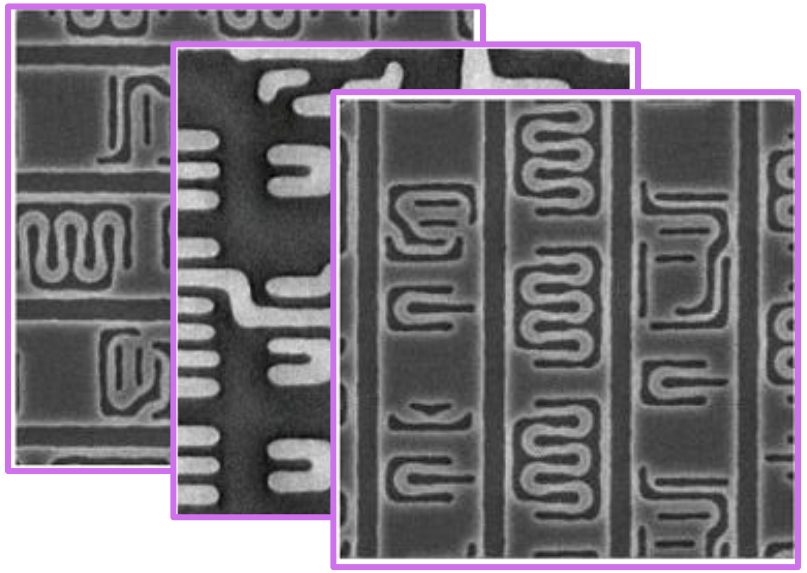
- Manufactured circuit very different from design due to quantum effects
- How do we get a good reference to identify defects?
  - Circuit design looks too different!
  - SEM (Scanning Electron Microscope) too slow to generate ground truth for every design / layer / patch

Train a GAN to Generate Reference Images

Circuit Designs *(d)*



SEM Images ("truth")

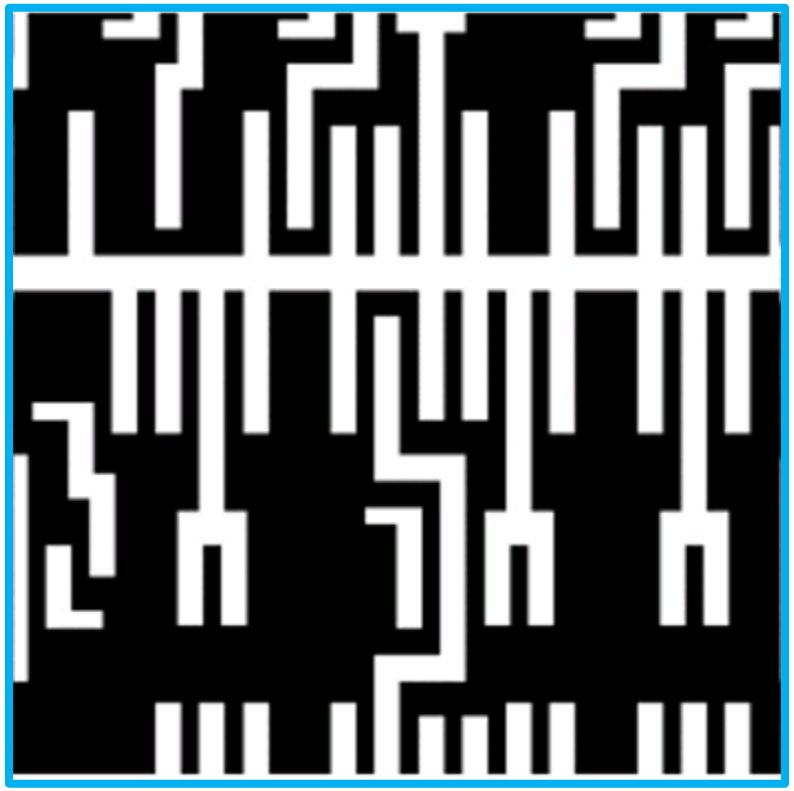


GAN Training

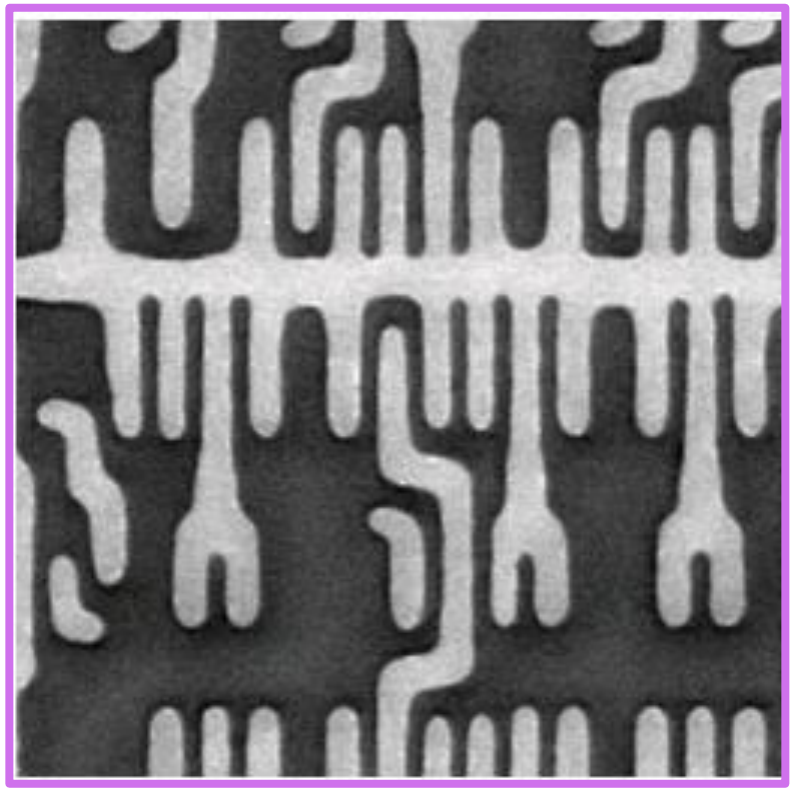
Generator *G(d)*

Generated Images Are Very Close to Ground Truth

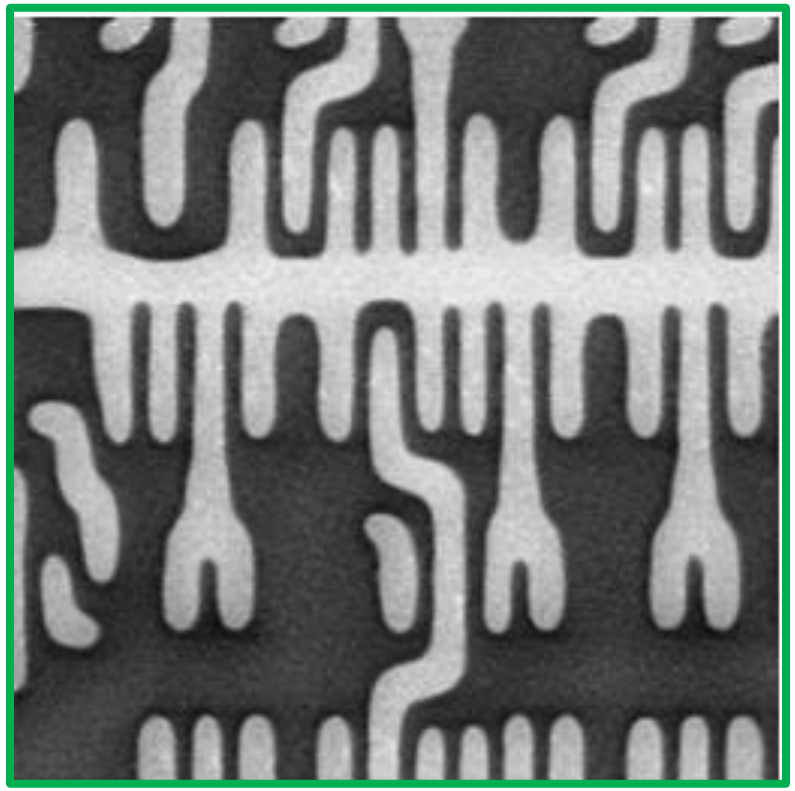
Subtle variations in patterns and imaging artifacts are faithfully reproduced



Circuit Design (Input, d)

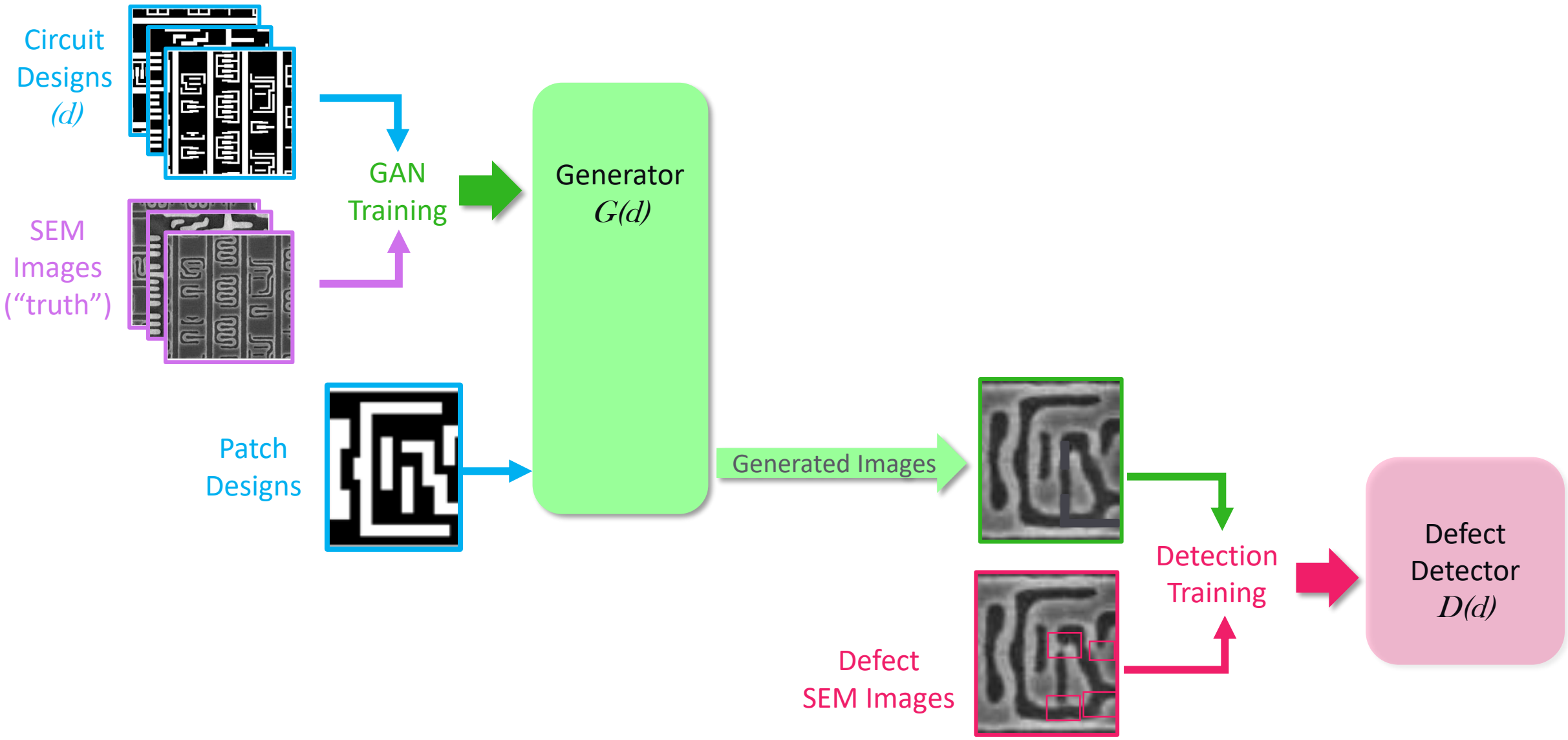


SEM Image (Ground Truth)



Generated Image, G(d)

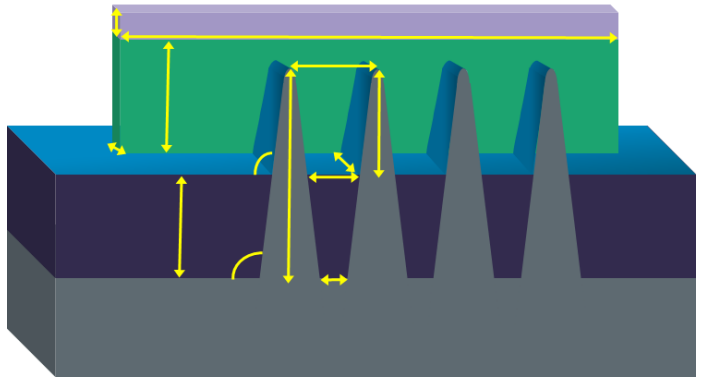
Generated Images Can Help Identify Defects



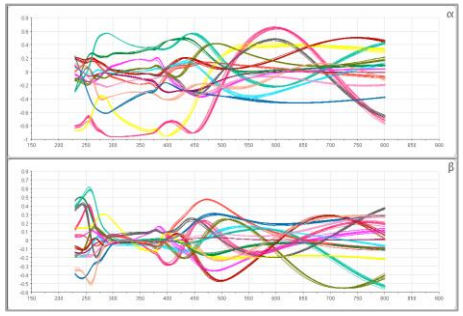
Metrology Using Machine Learning

Sample of Active DL R&D at KLA

Challenge: Measure Critical Dimensions to Track Process Variation

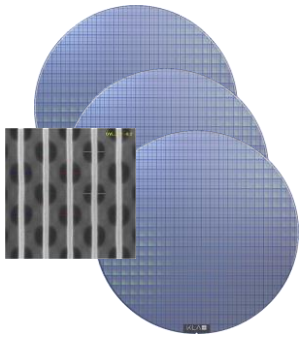


Parameterized Model



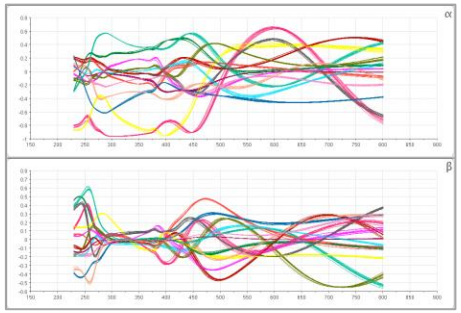
Simulated Spectra (S_{model})

Wafers



Metrology System

Spectra (S_{meas})

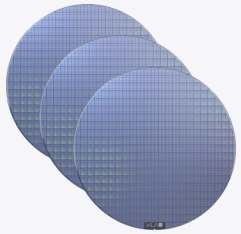


Regression Engine

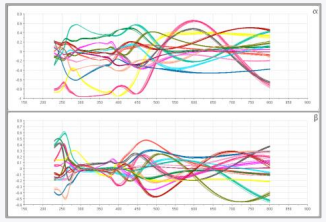
Parameters

Model-less vs. Model-based ML Recipe Development

Model-less Recipe Development



Wafers

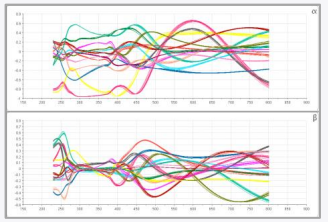


Reference

(CDSEM, TEM, e-Test…)

Trained ML Model

Model-based Recipe Development



Reference

Geometry (CD, HT, …)

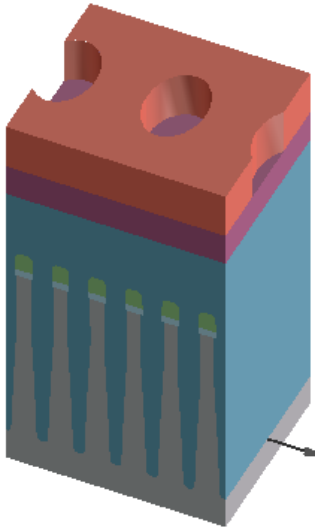
Trained ML Model

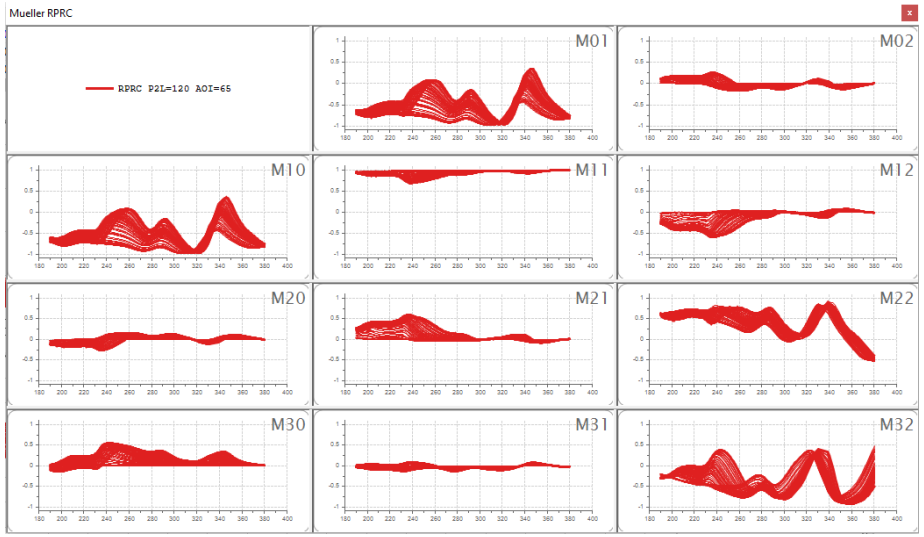
- Physics-assisted Machine Learning (ML) is part of our product portfolio for optical metrology tools
- It provides both model-less ML and enhanced model-based ML for robustness and matching

Challenges in Metrology

- Low sensitivity and high correlation parameters → challenging measurements

- Information in signals < information in model. DOF(Signal) < DOF(Model). How do we solve this?



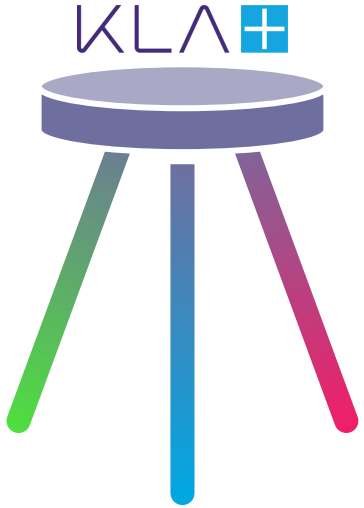


Challenging Issues Compared to Other Industry Applications

- Insufficient labeled (referenced) data
- Training samples may not cover large process variation (lack of training sample size)
- How to judge the quality of trained ML recipe for monitoring recipe robustness/process change without knowing ground truth?
- Multiple specs need to be achieved according to chip manufacturers' requirements (error control)
- Reference uncertainty

High Performance Computing: Making Rubber Meet the Road

KLA's Computation Stack



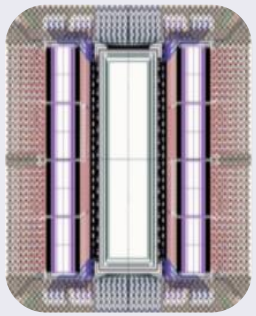
Optics

- Optical: 200-1000 nm
- SEM: 1 nm

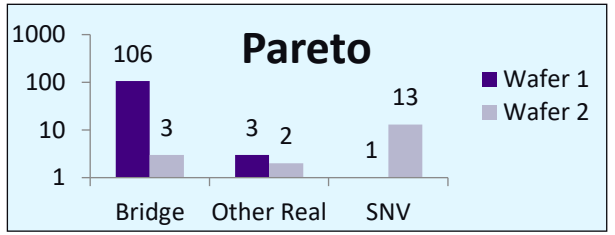
Image | Data Processing

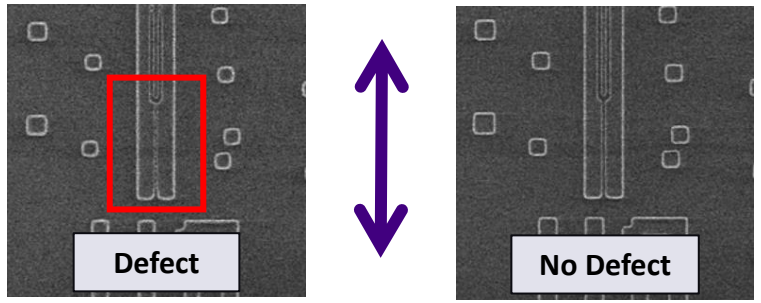
- Computing stack

High Speed Sensors

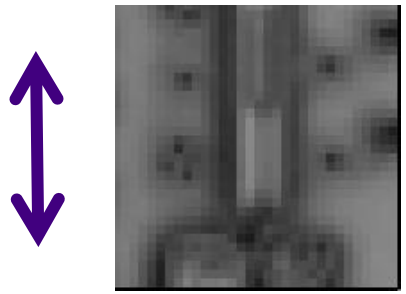


- 1-50 GB / sec





SEM Verification



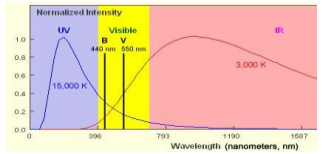
Optical Detection

CPU
Structured Data
GPU

GPU & CPU-SIMD
Bulk processing

FPGA
Near Real Time

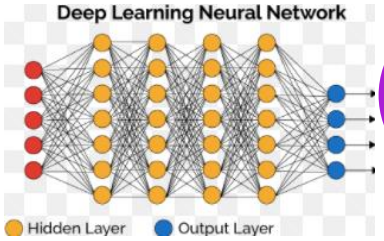
IMC Stack



ASIC/FPGA
Real Time

Raw Input

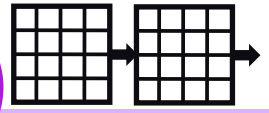
The Upside of moving Computations to GPUs

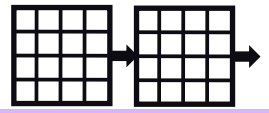


Deep Learning Networks

5-15X

GPU





Traditional Image Processing

3-6X

GPU

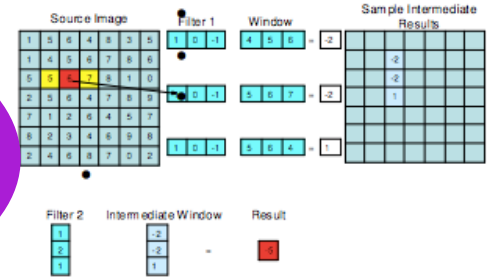
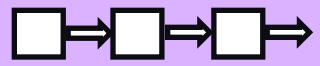


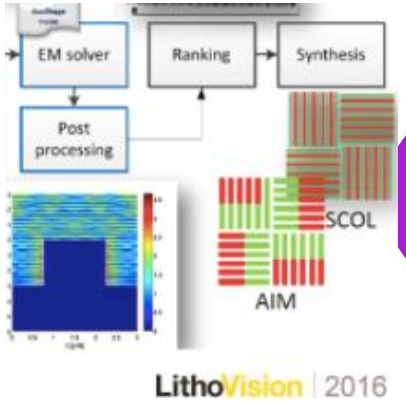
Figure 3. Separable filter convolution.

Traditional CPU Algos

GPU



GPU

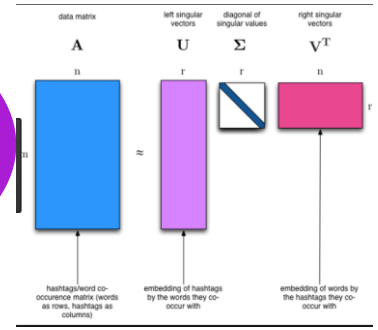


EM Solvers, 3D Tomography

2-4X

PCA, SVD, 2D-Histograms Small Matrix Solvers

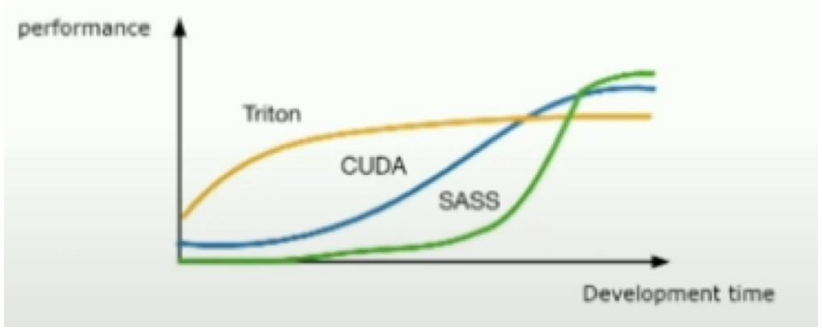
2-4X



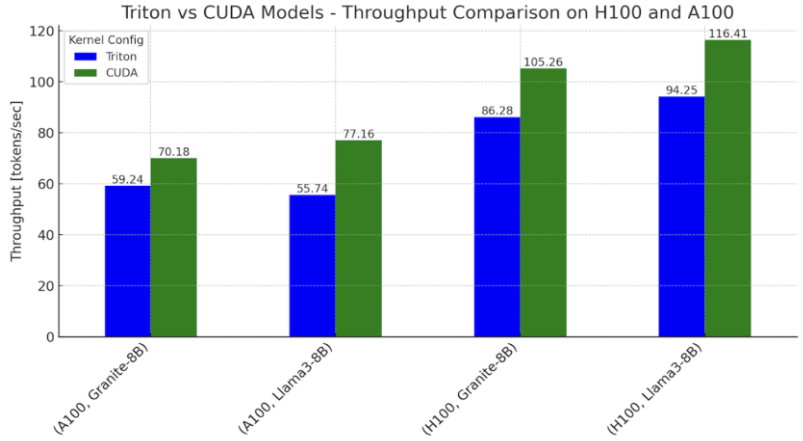
Moving to a Future Vendor-Agnostic SW Stack

- Performance portability is an important aspect to consider for future hardware
- OpenAI's Triton programming language has a lot of recent industry momentum
  - Python-like programming abstraction for device kernels with CUDA-like performance

[1]



[2]

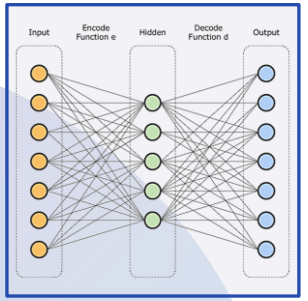


- Leveraging Compilers based on MLIR framework may deliver the holy-grail promise!
  - Progressive lowering enables data-science abstraction to reach perf of domain-specific hand-tuning

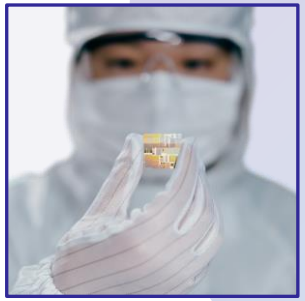
[1] Introducing Triton: Open-Source GPU Programming for neural networks https://openai.com/index/triton/
[2] CUDA-free inference for LLMs - https://pytorch.org/blog/cuda-free-inference-for-llms/

Our circle of AI & HPC

High Performance Computing



AI



Semiconductor Chips

Inspection & Metrology

Chip Manufacturing

In Conclusion

- Semiconductors are becoming an even more critical part of the global economy
- KLA's semi inspection & metrology tools enable continued scaling of Moore's law
- Inspection & Metrology requires cutting edge AI + HPC technologies to keep progressing

Why Join Us?

INVESTING IN INNOVATION

- We are committed to solving the most daunting technical challenges through innovation.
- We make large investments into research and development.

Thank you!